Report DDC-TR-69-1 AD-696 200

MACHINE-AIDED INDEXING

Paul H. Klingbiel
Directorate of Development

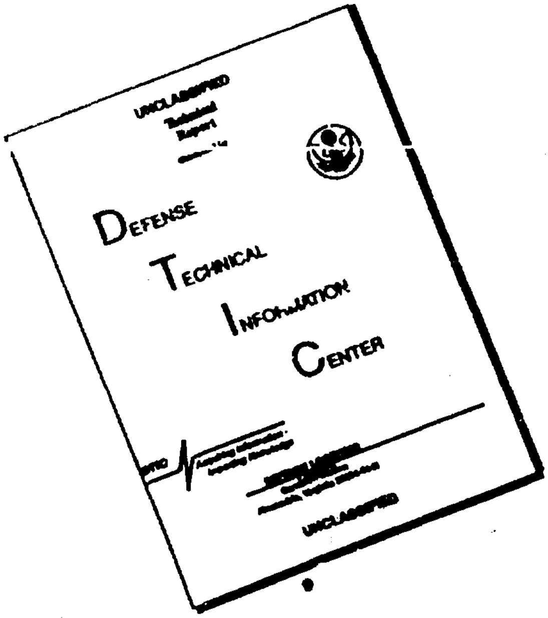
June 1969

Technical Progress Report for Period January 1967 - June 1969

DEFENSE DOCUMENTATION CENTER
Defense Supply Agency
Cameron Station
Alexandria, Virginia 22314

This document has been approved for public release and sale; its distribution is unlimited.

26



DEFENSE SUPPLY AGENCY
DEFENSE DOCUMENTATION CENTER
CAMERON STATION
ALEXANDRIA, VIRGINIA 22314

PREFACE

The Defense Documentation Center (DDC) regularly adds between 45,000 and 50,000 technical documents to its collection in a calendar year. Although many of these documents have been indexed by free language keywords at the source, DDC or its contractor - The Clearinghouse for Federal Scientific and Technical Information (CFSTI) - regularly reindexes these documents. The basic vocabulary has been a DDC-published thesaurus of about 7,400 authorized terms, an unpublished, classified listing of about 7,500 identifiers (military nomenclature, project names, etc.), and a growing list, also classified, of open-ended terms. The latter category, now in excess of 100,000 items, represents free language indexing.

Another collection, the Work Unit Information System (WUIS) (DD 1498), consisting of under 40,000 accessions, is also indexed by DDC with the vocabularies mentioned above.

DDC has, as an integral part of its mission, the responsibility for the development of new techniques for processing of technical information. The agency, therefore, attempts to maintain familiarity with the state of the art. In the area of indexing, particularly as that function might be either supplemented or taken over by a computer, DDC is familiar with the state of the art as represented by automatic indexing: A State-of-the-Art Report, M.E. Stevens, NBS Monography 91, 1965, and Progress and Prospects in Mechanized Indexing, M.E. Stevens, unpublished.

In general, there seems to be essentially two approaches to the problem of automatic indexing: (1) statistical analysis, or (2) syntactic analysis. The statistical technique requires fairly extensive stretches of text. DDC is constrained to work with titles and abstracts. For the DD 1498 collection this usually means less than 150 words per accession. For the technical report collection the stretches of text average about 200 words per accession. In addition, any machine indexing technique must compete with manual indexing costs to be of serious interest in the DDC production environment. Consequently, running time is extremely important. Running times would probably be prohibitively high for statistically-based indexing techniques. (Single-word indexing is not used by DDC; therefore, statistically-based systems would be required to generate word pairs, triples, etc.)

Complete syntactic analyses of sentences is not really within the state of the art if essentially one analysis per sentence is required. The assignment of multiple syntactic categories to single words guarantees ambiguity that cannot be automatically disambiguated by any algorithm known tothe author.

About 18 months ago the author introduced a technique for Machine-Aided Indexing (MAI) to the DDC staff as a viable approach in a production atmosphere. The indexing process does not depend upon a statistical analysis of the text or a simple kill list. Linguistic techniques are used, but complete syntactic analysis of sentences by computer are not required.

Simply stated, individual words are read into a computer and are either held for further consideration or eliminated from further processing. Lexical items such as commas, periods, and special symbols are recognized. The output is a list of candidate index terms and a screened exception list of terms and phrases for human review. Eventually the list of candidate terms will enter an Integrated Language Data Base that will have the capability of posting terms directly to the data base, switching synonyms to postable terms, or outputting nonrecognized terms for technical consideration.

The computer programs for the MAI System were written primarily in SLEUTH I, the assembly language for the UNIVAC 1107 (EXEC I). Some peripheral statistical programs (designed to compile information about the processes involved in text manipulation) and those programs used for the Language Data Base were written in COBOL for the UNIVAC 1108 (EXEC 8). Eventually, all programs for the system will be written for the UNIVAC 1108 running under EXEC 8.

The individual chapters of this paper follow the logic of the MAI System itself. The first chapter gives an overview of the entire process, and the succeeding chapters present a step-by-step account of the indexing procedure.

Component parts of the system are given in the upper case the first time they are mentioned together with an explanation of the use of that component. Thereafter, components are identified by initial capitalization.

Progress toward the goal of a system truly competitive with human indexing in cost, time, and comprehensiveness has been a team effort.

Prepared By:

Paul H. Klingbiel
PAUL H. KLINGBIEL
Directorate of Development

Approved By:

Herman W. Miles
HERMAN W. MILES
Director, Directorate of
Development

TABLE OF CONTENTS

THE LOGIC OF MACHINE-AIDED INDEXING

1. A lexical item is read into the computer and matched against a DISPOSITION DICTIONARY.

2. The Disposition Dictionary of lexical items carries a relative address for the single computer subroutine (macro) that controls the disposition of that item. The pilot model contains 13 such macros.

3. The disposition macros perform the following actions:

a. Hold a word in TEMPORARY STORAGE for future disposition.

b. Eliminate a word from all future consideration.

c. Print a word or group of words on an ERROR LISTING for technical editorial action.

d. Print a word or group of words on an INDEX TERM LIST.

4. A word held in Temporary Storage has its syntactic type - six types are recognized in the pilot model - stored in a secondary Temporary Storage location called the FORMAT REGISTER. Syntactic types are determined when a word is placed in the Disposition Dictionary. As successive words are stored in Temporary Storage, their syntactic types are recorded so that a syntactic formula is built up in the Format Register.

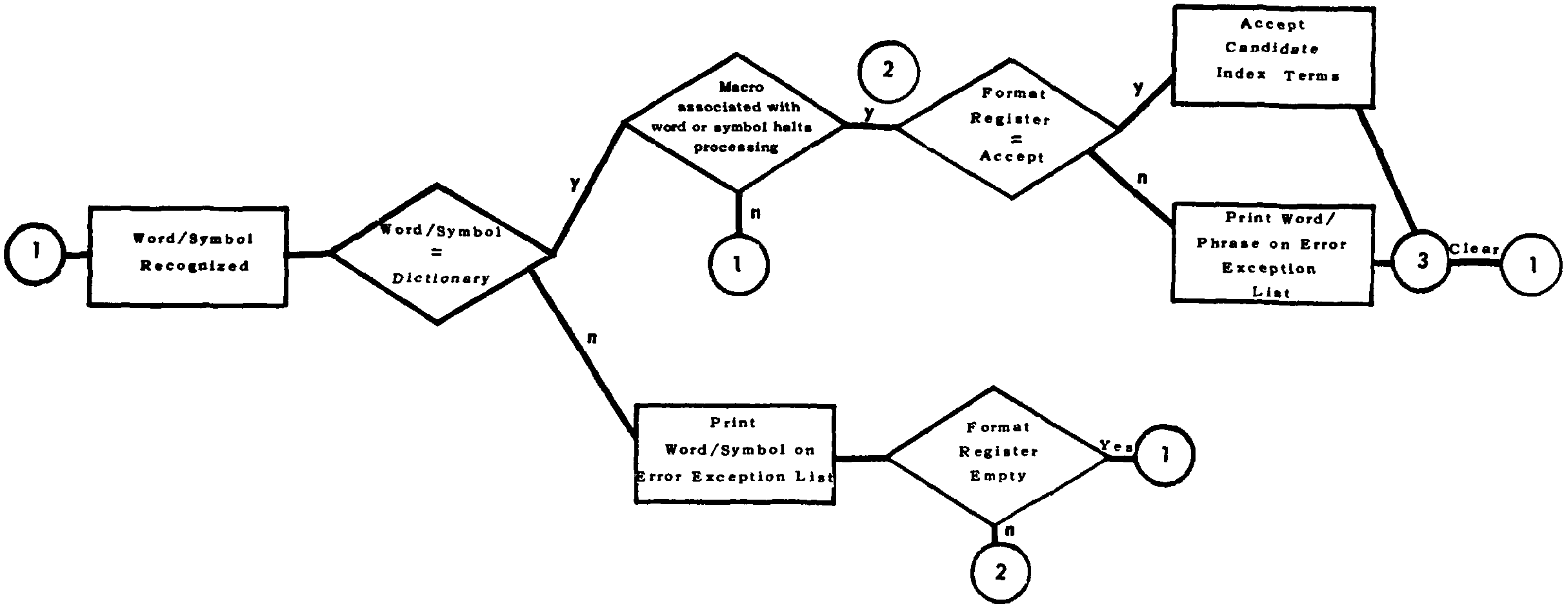
5. Eventually, a macro is called that prevents the addition of new words to Temporary Storage until the word or words already held there are moved. The effect of such a macro is to match the syntactic formula of the Format Register against the Format Dictionary of canonical formulas. This matching process has one of two results:

a. A match is made. The contents of Temporary Storage are printed as candidate index terms on the Index Term List (such a term may consist of more than one word). Both the Format Register and Temporary Storage are cleared and the indexing process proceeds by reading a new word in for matching against the Disposition Dictionary.

b. No match is made. The contents of Temporary Storage are printed out on the Error Listing for technical editorial review or, under certain conditions, the contents are deleted as being without further value. The Format Register and the Temporary Storage are cleared and the indexing process proceeds.

6. The logic of the system is briefly illustrated on the next page.

SYSTEM LOGIC CHART


1
Word/Symbol Recognized
Word/Symbol = Dictionary
y
n
Macro associated with word or symbol halts processing
n
1
y
2
Format Register = Accept
y
n
Accept Candidate Index Terms
Print Word/ Phrase on Error Exception List
3
Clear
1
Print Word/Symbol on Error Exception List
Format Register Empty
Yes
1
n
2

THE DISPOSITION DICTIONARY

When a lexical item is read into the computer, it is matched against a Disposition Dictionary. This dictionary consists of a two-element table of the following form:

Lexical item/disposition macro relative address.

The 13 entries below illustrate the range of syntactic types recognized by the current Disposition Dictionary:

The/Macro 1
Electric/Macro 2
Document/Macro 3
and/Macro 4
(Special Symbols)/Macro 5
Alloy/Macro 6
(Mismatch)/Macro 7
Of/Macro 8
(End of Field)/Macro 9
Or/Macro 10
Comma/Macro 11
Other/Macro 12
Space Hyphen Space/Macro 13

These macros can be discussed in two groups: Macros 1, 5, 7, 9, 11, and 13; Macros 2, 3, 4, 6, 8, 10, and 12.

Group I - Housekeeping Macros

Macro 1: Serves in part as a kill list. That is, words assigned to this category are eliminated from the indexing process. However, because of the conditional nature of the indexing macros, a simple exception list is not enough. Each time an item is deleted because of macro 1, the condition of the Format Register is checked. If the register is not empty, a complex set of conditions is tested to determine the nature of the contents in Temporary Storage. Only after this procedure has been followed and some disposition has been made of the contents of Temporary Storage can the indexing cycle be continued.

Macro 5: Disposes of the special symbols "apostrophe," "virgule," and the "closed" hyphen as in "two-story house." The macro instruction deletes those special symbols and allows the indexing procedure to advance. "Non" is recognized as a word and is marked as an adjective.

Macro 7: Triggers a printout for technical review when a word is received that is not in the Disposition Dictionary. This macro stops the indexing process until the contents of the Format Register are checked to see if useful terms have accumulated. After such a check all storage locations are cleared, and the indexing cycle resumes.

Macro 9: Signals when a complete computer field has been processed and moves the read function to the next field to be indexed after all registers have been properly cleared. (At present four fields of the DD 1498 are scanned for index terms: the title, objective, progress, and future plans.)

Macro 11: Stops the index process to check Temporary Storage for suitable index terms. (Index terms do not cross comma boundaries, except for the case of a sequence of adjectives, so the presence of a comma is used to check for useful index terms.)

Macro 13: Processes an orthographic idiosyncrasy; the horizontal line in "half-life conditions" must be processed differently than the horizontal lines in "injurious radiation - internal to the equipment - is screened."

Group II - Index Term Selectors

Macro 2: Places the lexical item in Temporary Storage and places an A (for adjective) in the Format Register.

Macro 3: Places the lexical item in Temporary Storage and places an N (for noun) in the Format Register.

Macro 4: Controls the disposition of "and." If Temporary Storage is empty, "and" is deleted and the next word is read into the computer. If Temporary Storage is not empty, "and" is placed in Temporary Storage and a "+" is placed in the Format Register.

Macro 6: Places the lexical item in storage and places a Z (for members of that class of nouns which cannot occur in isolation) in the Format Register.

Macro 8: Controls the disposition of the preposition "of." If the Format Register is empty, "of" is deleted and the next word is read in. If the Format Register is not empty, a complex set of conditions is checked to determine how the indexing process is to proceed.

Macro 10: Controls the disposition of the contents, if any, of the Format Register when "or" occurs in the text.

Macro 12: Controls the disposition of the contents, if any, of the Format Register when "other" occurs in the text.

BLANK PAGE

THE VARIETIES OF TEMPORARY STORAGE

Since the technique for MAI under discussion depends upon neither a statistical analysis nor a simple kill list, provision must be made to store information until enough data has been accumulated to render a decision as to whether or not an indexable word or phrase has been obtained. There are two varieties of Temporary Storage: the first variety accumulates the actual alpha representation of the index term possibilities; the second variety mirrors the alpha content of Temporary Storage by syntactic code. The second Temporary Storage device is called the Format Register; an abstraction of lexical items in terms of the next higher grammatical category is stored there.

The conditional nature of the decisions required by this kind of MAI was motivated by two factors: statistics obtained from human indexing and the importance of context. As examples of the statistical data that motivated the conditional approach, consider the index term frequencies of the following single terms taken from the AD collection at DDC:

TERM	NUMBER OF PCSTINGS
Design	77,828
Tests	51,881
Temperature	29,907
Measurement	38,154

These statistics, as of 30 June 1969, represent frequency of use by indexer in a collection of 580,000 documents. There is no way of knowing to what extent these figures represent textual frequency and to what extent they represent indexer idiosyncracy. The statistics do indicate that from a retrieval standpoint such single words in isolation carry little selectivity.

All of these words are of a general conceptual nature and are much more meaningful in combination:

TERM	NUMBER OF POSTINGS
Body Temperature	750
Desert Tests	710
High Temperature Alloys	6,922
Phase Measurement	618
Radiation Measurement Systems	1,809
Salt Spray Tests	1,629
Surface Temperatures	1,288
Temperature Coefficient of Reactivity	22
Temperature Sensitive Elements	405

These statistics are also for the AD collection as of 30 June 1969.

Having noted the desirable specificity when general terms are used in context, a study of context itself becomes important. In the list above, "temperature" plays several syntactic roles. In "body temperature" and "surface temperatures," "temperatures" is a noun. In the other four cases "temperature" functions adjectively. Analysis of this kind of data led to two conclusions: first, there should be a class of nouns that would be considered as index terms only when they occurred in combination with other terms; second, the usual impasse of a given lexical item functioning in two or more syntactic ways could be given an ad hoc solution that would eliminate the ambiguity.

The syntactic types of interest to the indexing function are stored in a permanent Format Dictionary. Syntactic formats are built up in the Format Register - the secondary Temporary Storage location - and are matched against the Format Dictionary on appropriate occasions. Matches between the syntactic formula held in Temporary Storage in the Format Register and the canonical formulas in the Format Dictionary become index terms; mismatches are printed out for scrutiny.

SYNTACTIC TYPE

The number of syntactic types employed by the DDC MAI system depends upon the definition of syntactic type. If one takes the existence of a unique macro as an indication of syntactic diversity, there are 13 syntactic types, or parts of speech. If, on the other hand, one is interested in just those parts of speech that constitute index terms or elements of index terms, there are six syntactic types.

The six possible components of index terms are: (1) N - class of nouns each of whose members is acceptable as a free form; (2) A - class of adjectives that can function only in the role of modifier; (3) Z - class of nouns each of whose members is acceptable only in combination with an N, or an A, or another Z, such as NZ, ZN, AZ, or ZZ, and of course strings of three or more such as AZN; (4) + - the word "and"; (5) P - the word "of"; and (6) C - the word "or."

The brief discussion of statistics and context given in the previous chapter can now be expanded. The fact that "temperature" tends to occur with high frequency and tends also to be nondiscriminating in isolation suggests that "temperature" be considered a "Z." The recognition that "temperature" can function adjectively does not preclude assigning a "Z" to the term - quite the reverse; the fact that "temperature" is desired only in combination strengthens the argument. From an indexing standpoint "low temperature alloys" is as logically represented by the syntactic formula AZZ as by AAZ. Moreover, "body temperature" requires either an NZ or a ZZ combination since "temperature" does occur in a noun head position.

Other investigators will raise serious questions as to whether the assignment of a unique syntactic type to a word is really feasible. Many examples of ambiguity can be produced that would seem to make such a unique assignment impossible. The approach the DDC investigation team has taken should be considered in terms of the following factors:

1. Only a subset of English is pertinent to indexing. For instance, verbs are never used as index points. The assignment of "N" to "programming" will lead to an acceptable situation for the string "linear programming theory." On the other hand "programming matrix calculations" will appear as an acceptable NNN form in "he was programming matrix calculations." However, in "he was programming and so was everyone else," "programming" will be picked up as an N with no harm done. The erroneous programming matrix calculations will be caught and rejected before posting on a file (see the chapter on Screens).

Adverbs ending in "ly" or "lly" are very rarely used as part of an index term. In those few cases that do exist, the word can easily be designated an adjective. The preponderant importance of noun and prepositional phrases in indexing has long been recognized. An early instance is Baxendale.[1] [2]

2. Nouns and adjectives can be distinguished as follows: An adjective is a word which never appears in isolation (or as a free form) as an index term. An adjective is always in a modifying, never a head position. This condition can be considered completely unambiguous since in natural scientific English the modifier precedes the noun head rather than following it. That is, the form suggested by "a woman scorned" and final parenthetic forms such as "conductivity (electrical)" are relatively rare and will print out on the Error List for technical review.

Nouns can appear either as heads of structures or modification or as modifiers. Plurals, which are a standard form of index terms, always appear either in isolation or in a head position. Whether such nouns are typed as N or as Z is a decision based largely on the noun's utility as a discriminating element in the data base for which the system is built. Singular nouns may appear in isolation or in a modifying position. Most such nouns will be categorized as Z: the decision is based on a study of occurrence through some such means as a permuted list.[3]

[1] Baxendale, P.B. "An Empirical Model for Computer Indexing" in "Machine Indexing," American University, 1962, pp. 207-218.

[2] Baxendale, P.B. "Man-Computer Indexing: Functions, Goals, and Realizations," in "Joint Man-Computer Indexing and Abstracting," MITRE SS-13, 1962, pp. 61-73.

[3] See the permuted listings in Thesaurus of Engineering and Scientific Terms, 1967, and the NASA Thesaurus, December 1967.

THE DATA BASE

The DDC experiment utilizes a subset of the Work Unit Information System (WUIS) (DD 1498) concerned with the Information Sciences. This set has been divided into 20 broad areas, and contains 2,447 resumes. Each work unit record has the following items of text, all of which are scanned by the computer indexing programs: Title, Field 12; Objective, Field 24; Approach, Field 25; and Progress, Field 26.

The initial experiment was conducted in Area IA, Data Compilation and Conventional Bibliography. This area contains 148 resumes. The four fields of interest contain 20,363 words of running text. A total of 3,350 word types were isolated, including punctuation marks. Punctuation is counted because of the role it plays in this MAI System. The disposition of the word types by macro was as follows:

135	Macro 1
282	Macro 2
298	Macro 3
1	Macro 4
3	Macro 5
696	Macro 6
1,930	Macro 7
1	Macro 8
Not a Vocabulary Item	Macro 9
1	Macro 10
1	Macro 11
1	Macro 12
1	Macro 13
3,350	

The Disposition Dictionary holds everything except Macro 7 terms, so that the effective dictionary size was 1,420.

The system was further refined by indexing Category IB, Scientific/ Technical Information and/or Data Centers. This area contains 328 work units, 43,433 running words, and 4,532 distinct word types. A merge of the 3,350 types from IA with the 4,532 types of IB resulted in the following macro disposition:

407	Macro 1
427	Macro 2
525	Macro 3
1	Macro 4
3	Macro 5
1,065	Macro 6
3,547	Macro 7
1	Macro 8
Not a Vocabulary Item	Macro 9
1	Macro 10
1	Macro 11
1	Macro 12
1	Macro 13
5,980	

The effective dictionary size was then 2,433.

Category IC, Information and/or Management System Studies, was then indexed. This category contains 233 resumes, 33,629 running words of text, and 3,968 word types. The merged word types (for 97,425 words of running text) resulted in the following macro distribution:

436	Macro 1
508	Macro 2
700	Macro 3
1	Macro 4
3	Macro 5
1,308	Macro 6
4,382	Macro 7
1	Macro 8
Not a Vocabulary Item	Macro 9
1	Macro 10
1	Macro 11
1	Macro 12
1	Macro 13
7,343	

The effective dictionary size now stands at 2,961. Dictionary growth is summarized in figure 1 on the following page.

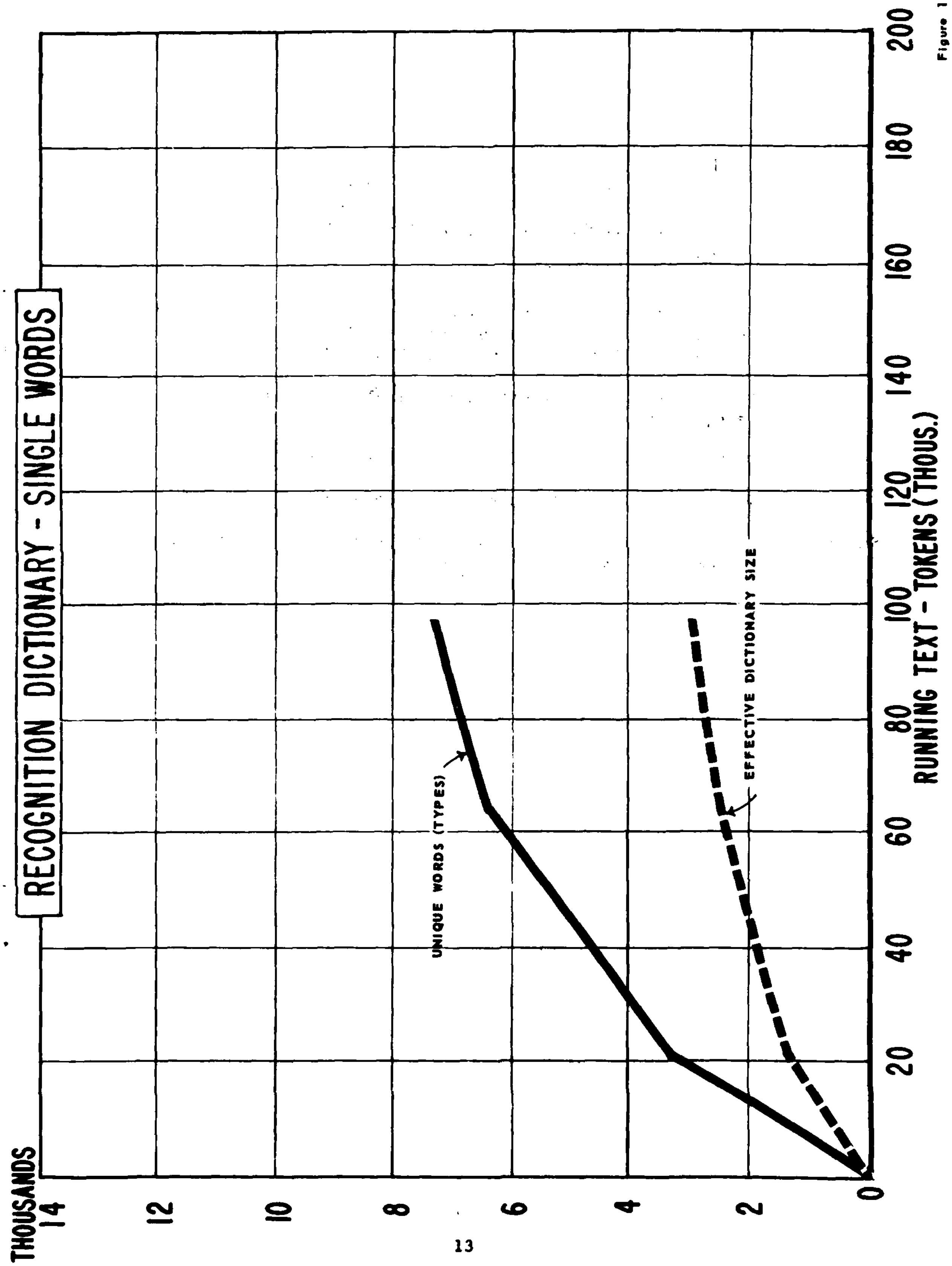

RECOGNITION DICTIONARY - SINGLE WORDS
THOUSANDS
14
12
10
8
6
4
2
0
20
40
60
80
100
120
140
160
180
200
RUNNING TEXT - TOKENS (THOUS.)
UNIQUE WORDS (TYPES)
EFFECTIVE DICTIONARY SIZE

Figure 1

The entire experimental data base is estimated to be about 350,000 running words. Dictionary size increase vs. total running words of text will be watched and statistics will be accumulated relative to syntactic type for index terms chosen.

Since this approach to MAI generates word combinations as well as single terms, the textual frequency of syntactic combinations will be investigated because a surprising feature of natural scientific text is the length of acceptable index phrases. Examples are:

1. Quasilinear uniformly elliptic partial differential equations and difference equations.

2. Problems of data management and data retrieval.

3. International data library and reference service.

Somehow such phrases must be reduced to an acceptable index term size of five words or less. Longer stretches are usually too specific.

SCREENS

MAI requires built-in contingency factors if anything like human indexer choices are to be made from text. A basic contingency factor is incorporated in the part-of-speech mechanism of stand-alone nouns vs. nouns requiring modification. However, candidate terms and phrases run a three-part hurdle before acceptance on the master file as bona fide retrieval points.

1. Words read in are either accepted for further analysis or rejected. Those words that are rejected are "killed" or printed out for technical review. Accepted words are held conditionally.

2. Words passed on for further analysis are stored and a syntactic formula is built up until the indexing process is halted by either a word reject, a conditional word such as "and," "or," or "of," or by punctuation. The accumulated syntactic formula is then checked with the format dictionary. A mismatch prints out the contents of Temporary Storage for technical review; a match transfers the candidate index terms to the third handle, the Integrated Language Data Base.

3. The Integrated Language Data Base is the final screen before posting. A match with a plural stand-alone noun is passed for posting; singulars of the same noun (these must be N's not Z's) are detected and posted on the plural form. A wide range of "use" references not involving plurals are also detected and posted on the preferred term. Long sequences of words of appropriate syntactic type will probably not match and will be displayed for technical review. Other screens are possible, but require investigation as to their utility.

BLANK PAGE

STATUS: 30 JUNE 1969

A pilot system has been developed that covers 709 DD 1498 resumes. The 97,425 words of running text are covered by a Disposition Dictionary of 2,961 terms, 13 macros, and 110 canonical syntactic formulas. With present programs the 709 resumes were machine-indexed in three minutes and forty seconds of CPU time.

Statistics are being collected on the frequency of occurrence of the various canonical syntactic formulas, the number of candidate index terms per document, and the distribution curves for index assignments. The frequency of occurrence, as seen by the Format Register, of the various canonical forms is illustrated by table 1, which lists the 25 most frequent forms in descending order.

Table 1

Rank	Type	Rank	Type
1	ZZ	14	ZPZZ
2	N	15	ZZ+Z
3	AZ	16	A+AZ
4	ZZZ	17	ZPN
5	AZZ	18	NZZ
6	Z+Z	19	AAZ
7	NZ	20	ZZZZ
8	ZPZ	21	Z+N
9	AN	22	N+Z
10	ZN	23	ZAZ
11	AZZZ	24	NN
12	ZPAZ	25	N+N
13	Z+Z		

Considering each acceptable format as a type, and its instances tokens, the 110 types generated 8,595 tokens. The top ranked "ZZ" is represented 1,659 times, the 25th ranked "N+N" is represented 50 times. Remember, "Z" in isolation is not a permissible form. It is startling to find AN ranked ninth and NN ranked in the twenty-fourth place. It would be reasonable to expect both of these types to occur more often and consequently to rank higher.

A typical DD 1498 is included (Page 19) which shows the text processed by the DDC indexing programs. The unedited candidate index terms that resulted are listed for comparison with the keywords supplied by the originator as well as the descriptors assigned by DDC analysts. A detailed comparison will be given in the next progress report. Additionally, the logic of the 13 macros is being optimized to further reduce running time. One of the ways to accomplish this is to investigate exhaustively the contexts within which certain words occur, such as: A, AND, OF, OR, BUT, OTHER, NON, and NOT.

Work is also progressing on the Integrated Language Data Base, which is the final screen for potential index terms before acceptance for posting on the Inverted File. That data base, in its initial form, will contain the majority of index terms and use references from TEST plus other terms and use references required by the MAI output. This component of the system will be more completely discussed in the next report.

RESEARCH AND TECHNOLOGY WORK UNIT SUMMARY				1. AGENCY ACCESSION* DN623796	2. DATE OF SUMMARY* 01 MAR 67		REPORT CONTROL SYMBOL T9
3. DATE PREV SUM'RY	4. KIND OF SUMMARY COMPLETED	5. SUMMARY SCTY* U	6. WORK SECURITY* U	7. REGRADING* N/A	8a. DISB'N INSTR'N	8b. SPECIFIC DATA-CONTRACTOR ACCESS ☒ YES □ NO	9. LEVEL OF SUM A. WORK UNIT

10. NO./CODES:*	PROGRAM ELEMENT	PROJECT NUMBER	TASK AREA NUMBER	WORK UNIT NUMBER
a. PRIMARY				
b. CONTRIBUTING				
c. CONTRIBUTING				

11. TITLE (Precede with Security Classification Code)*

(U) EXPLOSION DETECTION

12. SCIENTIFIC AND TECHNOLOGICAL AREAS*

015100 SEISMIC DET 010900 NUCLEAR EXPLOS

13. START DATE JANUARY 1965	14. ESTIMATED COMPLETION DATE MARCH 1966	15. FUNDING AGENCY DN	16. PERFORMANCE METHOD CONTRACT

17. CONTRACT/GRANT

a. DATES/EFFECTIVE: EXPIRATION:

b. NUMBER:*

c. TYPE: d. AMOUNT:

e. KIND OF AWARD:

18. RESOURCES ESTIMATE		a. PROFESSIONAL MAN YRS	b. FUNDS (In thousands)
FISCAL YEAR	PRECEDING 65	2.0	112
	CURRENT 66	1.0	

19. RESPONSIBLE DOD ORGANIZATION 265250 0900

NAME:* OFFICE OF NAVAL RESEARCH
ADDRESS:* WASHINGTON D.C. 20360

RESPONSIBLE INDIVIDUAL
NAME: WINCHESTER, J.W.
TELEPHONE: 202-OX-6-6967

20. PERFORMING ORGANIZATION 060100 2208

NAME:* BOLT, BERANEX & NEWMAN
ADDRESS:* 50 MOULTON ST. CAMBRIDGE, MASS
02138

PRINCIPAL INVESTIGATOR (Furnish SSAN if U.S. Academic Institution)
NAME:* MARILL, T
TELEPHONE:
SOCIAL SECURITY ACCOUNT NUMBER:

21. GENERAL USE

ASSOCIATE INVESTIGATORS
NAME:
NAME:

22. KEYWORDS (Precede EACH with Security Classification Code)

NUCLEAR DETECTION; SENSORS; SEISMIC SIGNALS

23. TECHNICAL OBJECTIVE.* 24. APPROACH. 25. PROGRESS (Furnish individual paragraphs identified by number. Precede text of each with Security Classification Code.)

24. (U) DETERMINE A SYSTEM FOR DATA HANDLING WHICH PERMITS A DECISION TO BE MADE FROM MANY SENSOR INPUTS ABOUT THE CLASSIFICATION OF AN EVENT AS AN EARTHQUAKE OR A NUCLEAR EXPLOSION. THE VISTA SYSTEM (VISUAL STATISTICAL ANALYSIS) IS TO BE ADAPTED TO THE USE OF NUCLEAR EXPLOSION DETECTION USING SEISMIC SIGNALS AS INPUT. THIS WORK UNIT IS A PORTION OF THE VELA UNIFORM TASK OF NAVY INTEREST. CLASSIFICATION OF EVENTS AS NATURAL OR EXPLOSIVE REMAINS A MAJOR DIFFICULTY IN THE DEVELOPMENT OF A SURVEILLANCE SYSTEM.

26. (U) SAMPLE DATA HAVE BEEN SELECTED WHICH INCLUDE EARTHQUAKE AND NUCLEAR EVENTS AND THESE DATA WILL BE SUBJECTED TO ANALYSIS TO OBTAIN POSSIBLE CRITERIA FOR CATEGORIZING THE EVENTS.

RETRIEVAL TERMS ASSIGNED BY DDC: EARTHQUAKES; DETECTORS; CLASSIFICATION; SEISMIC WAVES; NUCLEAR EXPLOSIONS; STATISTICAL ANALYSIS; VISTA (VISUAL STATISTICAL ANALYSIS); IST.

*Available to contractors upon originator's approval

CANDIDATE INDEX TERMS FOR SPECIMEN 1498

Type	Terms
ZZ	EXPLOSION DETECTION
ZZ	DATA HANDLING
AZ	NUCLEAR EXPLOSION
NZ	VISTA SYSTEM
AAZ	VISUAL STATISTICAL ANALYSIS
ZPAZZ	USE OF NUCLEAR EXPLOSION DETECTION
AZ	SEISMIC SIGNALS
N	VELA
N	NAVY
NZ	SURVEILLANCE SYSTEM

EXCEPTION LIST

Terms	Diagnostic
PERMITS	MAC 7
INPUTS	MAC 7
EVENT	MAC 7
ADAPTED	MAC 7
UNIFORM	MAC 7
TASK	MAC 7
CLASSIFICATION	NON-MATCH
EVENTS	MAC 7
NATURAL	ADJ
REMAINS	MAC 7
SAMPLE	MAC 7
EVENTS	MAC 7
SUBJECTED	MAC 7
CATEGORIZING	MAC 7
EVENTS	MAC 7

PLANS FOR CALENDAR YEAR 1969

The Format Dictionary is being replaced with a recursive right linear grammar which, in Greibach normal form, will accept canonical forms of any length. Only 27 rules are required. This system is being programmed, and running times will be compared with the original system.

Right linear grammars have also been written that can be used to recognize all well-formed authorized AN numbers as well as well-formed and authorized contract numbers. These are also being programmed and tested for efficiency.

Several thousand DD 1498 resumes are being indexed to build a file that will permit parallel searching of live requests to test the adequacy of MAI terms for retrieval.

The Integrated Language Data Base is being enlarged both in size and in capability. If the grammars prove to be efficient devices in terms of running time, they will be incorporated into the data base for increased sophistication.

Cost/benefit statistics will be collected for comparison of MAI with manual methods.

A status report will be prepared.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Defense Documentation Center

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

b. GROUP
N/A

3. REPORT TITLE
Machine-Aided Indexing

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*
PAUL H. KLINGBIEL

6. REPORT DATE
30 June 1969

7a. TOTAL NO. OF PAGES
28

7b. NO. OF REFS
Five (5)

8a. CONTRACT OR GRANT NO.

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

A partial syntactic analysis is used to detect words and phrases in contexts which make them useful for indexing purposes. For instance, the word "abstracted" is useful only when it functions as an adjective. A total of 97,425 words of text have been run through the index programs in three minutes and forty seconds on the UNIVAC 1108 under EXEC I. The output is being analyzed to detect synonyms and to compare the machine-produced index terms with manual indexing assignments. At least 500,000 words of text will be processed to obtain statistics to determine whether the system is competitive with manual indexing.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Machine Aided Indexing						
Manual Indexing						
Parts of Speech						
Index Programs						
Recognition Dictionary						
Candidate Index Terms						
Dictionary						
Words						
Synonyms						